Solutions to Reinforcement Learning by Sutton Exercise 9.6

Simon Haastert

November 2022

1 Exercise 9.6

It is given:

$$\tau = 1 \tag{1}$$
$$\mathbb{E}[\mathbf{x}^\top \mathbf{x}] = \mathbf{x}\left(S_t\right)^\top \mathbf{x}\left(S_t\right) \tag{2}$$
$$\begin{aligned} \alpha &= \left(\tau \mathbb{E}\left[\mathbf{x}^\top \mathbf{x}\right]\right)^{-1} \\ &= \left(1 \cdot \mathbf{x}\left(S_t\right)^\top \mathbf{x}\left(S_t\right)\right)^{-1} \end{aligned} \tag{3}$$
$$\mathbf{w}^{t+1} = \mathbf{w}^t + \alpha \left[U_t - \hat{v}(S_t, \mathbf{w}^t)\right] \nabla \hat{v}(S_t, \mathbf{w}^t) \tag{4}$$

We apply linear function approximation. Thus:

$$\hat{v}(S_t, \mathbf{w}^t) = (\mathbf{w}^t)^\top \mathbf{x}(S_t) = \mathbf{x}(S_t)^\top \mathbf{w}^t \tag{5}$$
$$\nabla \hat{v}(S_t, \mathbf{w}^t) = \mathbf{x}(S_t) \tag{6}$$

Plugging in (5) and (6), we get:

$$\mathbf{w}^{t+1} = \mathbf{w}^t + \alpha \left[U_t - \mathbf{x}(S_t)^\top \mathbf{w}^t\right] \mathbf{x}(S_t) \tag{7}$$

Now we insert (3) for α in (7):

$$\begin{aligned} \mathbf{w}^{t+1} &= \mathbf{w}^t + \left(\mathbf{x}\left(S_t\right)^\top \mathbf{x}\left(S_t\right)\right)^{-1} \left[U_t - \mathbf{x}(S_t)^\top \mathbf{w}^t\right] \mathbf{x}(S_t) \\ &= \mathbf{w}^t + \left[U_t - \mathbf{x}(S_t)^\top \mathbf{w}^t\right] \mathbf{x}(S_t) \left(\mathbf{x}\left(S_t\right)^\top \mathbf{x}\left(S_t\right)\right)^{-1} \\ &= \mathbf{w}^t + \left[U_t - \mathbf{x}(S_t)^\top \mathbf{w}^t\right] \mathbf{x}(S_t) \mathbf{x}\left(S_t\right)^{-1} \left(\mathbf{x}\left(S_t\right)^\top\right)^{-1} \\ &= \mathbf{w}^t + \left[U_t - \mathbf{x}(S_t)^\top \mathbf{w}^t\right] \left(\mathbf{x}\left(S_t\right)^\top\right)^{-1} \end{aligned} \tag{8}$$

Because $\left[U_t - \mathbf{x}(S_t)^\top \mathbf{w}^t\right]$ is a scalar:

$$\begin{aligned}
\mathbf{w}^{t+1} &= \mathbf{w}^t + \left(\mathbf{x}\left(S_t\right)^\top\right)^{-1} \left[U_t - \mathbf{x}(S_t)^\top \mathbf{w}^t\right] \\
&= \mathbf{w}^t + \left(\mathbf{x}\left(S_t\right)^\top\right)^{-1} U_t - \left(\mathbf{x}\left(S_t\right)^\top\right)^{-1} \mathbf{x}(S_t)^\top \mathbf{w}^t \\
&= \mathbf{w}^t + \left(\mathbf{x}\left(S_t\right)^\top\right)^{-1} U_t - \mathbf{w}^t \\
&= \left(\mathbf{x}\left(S_t\right)^\top\right)^{-1} U_t
\end{aligned} \tag{9}$$

Now it is easy to see that the error is being reduced to zero in one update by plugging in the new weights in the next update step (see equation (7)):

$$\begin{aligned}
\mathbf{w}^{t+2} &= \mathbf{w}^{t+1} + \alpha \left[U_t - \mathbf{x}(S_t)^\top \mathbf{w}^{t+1}\right] \mathbf{x}(S_t) \\
&= \mathbf{w}^{t+1} + \alpha \left[U_t - \mathbf{x}(S_t)^\top \left(\mathbf{x}\left(S_t\right)^\top\right)^{-1} U_t\right] \mathbf{x}(S_t) \\
&= \mathbf{w}^{t+1} + \alpha \left[U_t - U_t\right] \mathbf{x}(S_t) \\
&= \mathbf{w}^{t+1}
\end{aligned} \tag{10}$$